Fast Computation of Statistical Analysis for Large Spatial Data Sets

著者	Hirano Toshihiro
内容記述	この博士論文は内容の要約のみの公開（または一部非公開）になっています
year	2018
その他のタイトル	大規模空間データに対する統計解析の高速計算
学位授与大学	筑波大学 (University of Tsukuba)
学位授与年度	2017
報告番号	12102甲第8511号
URL	http://hdl.handle.net/2241/00152228

博士(社会工学)論文概要

Fast Computation of Statistical Analysis for Large Spatial Data Sets
(大規模空間データに対する統計解析の高速計算)

システム情報工学研究科　社会工学専攻
社会工学学位プログラム

平野　敏弘

２０１８年　３月

Hirano, Toshihiro

Fast Computation of Statistical Analysis for Large Spatial Data Sets

Thesis directed by Prof. Morito Tsutsumi

This thesis discusses spatial statistics, that is, statistics for spatial data. Data with spatial information, such as latitude, longitude, and sampling district are called spatial data, and are collected through sensors, census, satellites, and environmental surveys.

Spatial data are divided into three categories based on the types of spatial information attached to them. The first type of spatial data has coordinates as the spatial information. For a model of this type of data, a stochastic process, which is called a random field, is adopted. The data on precipitation, air dose rate, and humidity are regarded as spatial data with coordinates, and mainly studied in geostatistics and environmental studies. The spatial variation is formulated as a covariance function in the random field and often depends on the Euclidean distance. The second category of spatial data is called areal data and is sampled in each district, city, and country. The test score, annual profit, and population in each field are regarded as areal data. Spatial econometrics focuses mainly on this second type of spatial data sets because of the economic data in them. The third category is point pattern data, which often captures the place of occurrence of crimes, traffic accidents, and other incidents. For example, the analysis of this type of spatial data set reveals the dangerous regions in a large territory through the estimation of an intensity function.

Advances in Global Navigation Satellite System (GNSS) and compact sensing devices make it easy to collect a large volume of spatial data with coordinates in the natural and social sciences. The sample size of this type of data, which is called large spatial data sets, can range from 5,000 to 150,000, and they generally consist of urban data (e.g., traffic flow, land price, and water demand) and environmental data (e.g., precipitation, air dose rates, and mineral content).

The statistical analysis such as model fitting and prediction for large spatial data sets

would help in the management of a smart city and creation of an evidence-based environment policy. However, it has been widely recognized that the conventional spatial statistical methods are impractically time-consuming for large spatial data sets. This makes it difficult for practitioners to utilize large spatial data sets for various applications.

This difficulty in utilization has encouraged the development of efficient statistical methods for large spatial data sets and has also led to the publication of many studies over the past decade. The key idea in the earlier efficient statistical methods is the approximation of the original statistical model by a computationally feasible statistical model. Consequently, fast computation is achieved at the cost of a slight decay in the accuracy of the estimation and prediction. Note that these studies have discovered some important asymptotic properties in spatial statistics and have contributed to the progress of spatial statistics.

In this thesis, we have addressed the computational burden in the statistical analysis of large spatial data sets. Our main contributions are to survey many earlier fast computation methods for large spatial data sets and propose two novel and efficient statistical methods.

This thesis is organized as follows. In Chapter 1, we introduce basics of spatial statistics. To begin with, we explain that the spatial data is modeled as the stochastic process, which is called the random field. The stationary random field has the covariance function which controls the spatial similarity. To explicitly incorporate this type of similarity into a statistical model with the covariance function is one of the characteristics of spatial statistics. Some typical covariance functions are introduced. The Matérn covariance function is, in particular, the most important one in spatial statistics because it can determine the smoothness of the random field. Then, in the random field, we summarize the conventional elementary spatial statistical methods, such as the maximum likelihood estimator (MLE), the kriging predictor, and the Bayesian estimation and prediction with the Markov chain Monte Carlo (MCMC) method.

In Chapter 2, we survey a large amount of earlier studies of efficient statistical methods for large spatial data sets. To begin with, we indicate that the cause of the computational

burden in the estimation and prediction for large spatial data sets is the inversion of the covariance matrix, which is included in the conventional spatial statistical methods mentioned in Chapter 1. The operation count for its direct computation is of order n^3 for sample size n. Hence, as the sample size increases, the computation becomes a formidable barrier in practice. To deal with this problem, many methods have been developed. These techniques are categorized into three types: a sparse approach, a low rank approach, and a spectral approach. We review the details of each approach.

In Chapter 3, we present a novel estimator that enables fast calculation of the coefficients in a spatial linear regression model for the integer lattice points in two-dimensional Euclidean space. When estimating coefficients in the linear regression model, a generalized least squares estimator (GLSE) is used if the covariance structures are known. However, the GLSE for large spatial data sets is computationally expensive because it involves the inversion of the covariance matrix of error terms for each observation. Therefore, we propose a pseudo best estimator (PBE) using spatial covariance structures approximated by separable covariance functions, that is, the product of the covariance functions of the causal autoregressive (AR) processes. The inversion of the resulting covariance matrix is easily calculated because the resulting covariance matrix is expressed as the Kronecker product of the covariance matrix of the two AR processes, and its inversion is identical with the the Kronecker product of the inversion of each covariance matrix of the AR processes. By extending the mathematical technique developed in the time series literature to the spatial case, the asymptotic covariance matrix of the proposed estimator is derived. Some simulations show that our proposed estimator outperforms the GLSE and a least squares estimator (LSE) in estimation accuracy or the calculation time, even if the true covariance function is isotropic. In particular, when the LSE is not asymptotically efficient, the PBE exhibits superior performance. Moreover, in empirical data analysis based on the soil moisture index, our proposed method supports the effectiveness in the estimation accuracy or the computation time.

In Chapter 4, we propose a modification of an existing fast computation technique,

which is called the linear projection, for large spatial data sets. The advantage of the linear projection approach is that it avoids the knot selection problem which the Gaussian predictive process faces with, while it cannot overcome the low approximation accuracy of the small-scale spatial variation. To deal with this problem, we developed the modified linear projection, where the approximation error of the linear projection is improved by using the covariance tapering. This proposed method is easy to implement and can capture both the large- and small-scale spatial variations effectively. To begin with, we show that the Kullback–Leibler divergence between the original probability density function and that of the modified linear projection has a sharper bound than the linear projection. Next, through simulated examples and empirical data analysis based on the air dose rate in the Chiba prefecture of eastern Japan via the MCMC method, we demonstrate that our proposed approach generally performs well in terms of computation time, estimation of model parameters, prediction at unobserved locations, and adequacy of the model regardless of the dependence magnitude of the spatial covariance functions when compared with some benchmark methods. Additionally, although the motivation to develop the modified linear projection was the need to improve the approximation of the original stationary covariance function in the linear projection, the empirical study has shown that it can also be used as a valid nonstationary covariance function model beyond just an approximation. This is why the modified linear projection approximates the original stationary covariance function by the nonstationary one.

A summary of the thesis and suggestions for future studies are presented in Chapter 5. We describe three tasks to provide a direction for future research. The first task is to further improve the modified linear projection proposed in Chapter 4. It is known that the modified linear projection approach does not perform well for the nonstationary random field and/or the spatial data with strong spatial covariance around the origin. This is why the covariance tapering, which is used to modify the linear projection, is generally a stationary and compactly supported correlation function. We expect to improve the estimation and

prediction accuracy of the original linear projection substantially by incorporating it into the multi-resolution approach. Second, since there is a rich literature on fast computation techniques for large spatial data sets in the univariate random field, it would be important to develop new fast computation techniques for the multivariate random field. Finally, the ultimate goal in this field is to deal with large multivariate spatio-temporal data sets efficiently. Since multivariate spatio-temporal data sets not only have the enormously large sample size but also are sequentially observed, it would be necessary to develop an online statistical technique for fast computation.